## How LightGBM is Used in the BERT + LightGBM Hybrid Model?

1. **BERT Extracts Semantic Meaning**
   - We take the product's item_name (e.g., *"Up & Go Breakfast Drink Vanilla 3x250ml"*)
   - Use **DistilBERT** to convert that name into a **768-dimensional vector**
   - This vector captures the **semantic context** (brand, size, type, flavor, etc.)

2. **PCA Reduces Dimensions**
   - 768 is too high-dimensional for LightGBM
   - So we apply PCA to reduce it to 100 dimensions
   - This keeps the most informative signals while reducing noise and complexity

3. **Structured Features are added**
   We combine the reduced BERT features with classic tabular data like:
   - item_price
   - unit_price
   - DiscountRate
   - PriceCapped
   - Week_num
   - was_on_special_last_week

4. **LightGBM trains on combined data**
   - We now have a single feature matrix:
     X = [structured features + PCA-reduced BERT embeddings]
     y = next_on_sale_week (multi-class label: Week 1–8)
   - LightGBM is trained as a multi-class classifier to predict the correct week

## Why Use LightGBM?

| **Reason** | **Benefit** |
|---|---|
| Handles tabular + numerical data well | Perfect for structured retail features |
| Very fast to train | Ideal when BERT embeddings are precomputed |
| Easy to interpret | Feature importance, tree plots |
| Supports calibration | Used with CalibratedClassifierCV for better probabilities |

**Summary:**

**BERT** handles the **semantic understanding** of product names.
**LightGBM** does the **actual week classification**, using both **semantic + structured signals**.

## Final Summary & Conclusion:

**Key Takeaways:**

| **Milestone** | **Accomplishment** |
|---|---|
| Dataset Generation | Realistic 8-week synthetic dataset based on original data |
| Quality Assurance | Advanced logic, statistical, and visual checks |
| Model Development | High-performing hybrid of BERT and LightGBM |
| Results | Reliable predictions on discount week likelihood |

**Limitations**

- Synthetic logic doesn't reflect **real market demand or inventory factors**
- BERT was not **fine-tuned** — potential for future accuracy improvements
- Lack of real transactional feedback